Data Cleaning with Open Refine

●●●

Alison Blaine, ablaine@ncsu.edu
Jennifer Garrett, jigarret@ncsu.edu
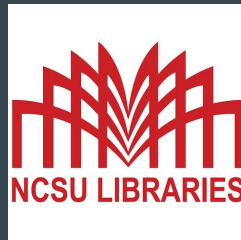
NCSU LIBRARIES

Data Visualization Services

lib.ncsu.edu/do/visualization

Data Visualization Process

- getting data
- getting to know the data
- setting goals
- determining if a visualization is needed
- deciding what to visualize
- cleaning
- visualizing
- potentially re-visualizing it with a different tool
- exporting/embedding

Data cleaning

Cleaning data means removing problematic characters, whitespaces, symbols, errors, missing elements and inconsistencies.

C
Raleigh
Raleigh, NC
Raleigh, N.C.
Raleigh, NC 27695
Raleigh, NC USA

Tools:

- Open Refine
- ASAP Utilities for Excel
- Data cleaning using code in R or SAS

Open Refine

- Formerly owned by Google, now maintained by independent developers.
- Free, open source tool that makes it easy to clean and transform large data sets.
- You can find problems in your data set without knowing about them in advance.
- Requires downloading the software in order to use (http://openrefine.org/download.html)
- GREL documentation: https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

Today's workshop materials

go.ncsu.edu/refine